Supplementary materials of *HESS-2020-43*

Replies to the comments of Reviewer #3

Interactive comment on "*A flexible two-stage approach for blending multiple satellite precipitation estimates and rain gauge observations: an experiment in the northeastern Tibetan Plateau*" by Yingzhao Ma et al.

Yingzhao Ma et al.: yzma@colostate.edu

Topic: More details of the two-stage blending (TSB) approach

3 The TSB algorithm

3.1 Overview

This algorithm aims at developing a multi-source data merging framework to provide the best-available gridded precipitation product with GR and SPE in the region of interest. Let $R(s,t)$ denote near-surface precipitation at the GR cell s and the t^{th} day. The original SPE and bias-corrected SPE are defined as $(Y_1(s,t), Y_2(s,t), \dots, Y_p(s,t)$ and $(Y_1'(s,t), Y_2'(s,t), \dots, Y_p'(s,t))$ at the same grid and time, respectively. For simplicity, they are separately replaced by R, $(Y_1, Y_2, \dots, Y_p)$, and $(Y_1', Y_2', \dots, Y_p')$. The subscript p implies the number of SPE in terms of its value at 4 in the following application, and PERCDR, 3B42V7, CMORPH and IMERG refer to Y_1, Y_2, Y_3, Y_4, respectively.

The diagram of the TSB method is shown in Figure 2. Stage 1 is designed to mitigate the bias of SPE based on the GR at the training sites with a Bayesian correction (BC) procedure, where the assumption of probabilistic distribution for GR conditional on each SPE is not limited to Gaussian prototype. Given complex terrain and 0.25° grid resolution, the topography is added as a covariate in the BC process. In the 2nd stage, a Bayesian weight (BW) model is used to merge the bias-corrected SPE. The BW model can exert benefit from bias-adjusted SPE with high performance and reduce poor impact from the ones with lower quality. It also produces blended SPE with predictive uncertainties. The details of the TSB algorithm are described in Sections 2.2 and 2.3, respectively.

3.2 Stage 1: Bias correction

In this stage, we perform on conditional modelling of GR on each SPE, i.e., on the probabilistic distribution $f(R)$ at the training sets to improve the accuracy of the original SPE. A flexible assumption (e.g., Lognormal, Gaussian, or Student's t distribution) for bias characteristics between GR and SPE is proposed. Given various SPE at different training sites, the specific probabilistic function is not limited to a certain distribution. For demonstration purposes, we herein apply the Student's *t* distribution, with its mean parameter expressed as a linear regression of the original SPE and terrain feature in the case. It is parameterized below:

$$R \sim Student(\nu_i, \mu_i, \sigma_i) \tag{1}$$

$$\mu_i = \alpha_i + \beta_i * Y_i + \gamma_i * Z \tag{2}$$

where ν_i is known as degree of freedom, μ_i and σ_i stand for sample mean and variance, respectively; the parameter μ_i is correlated with the intensity value of the i^{th} SPE (Y_i) and terrain feature (Z). To ignore data anomaly, the elevation feature in Eq. (2) is normalized with its value ranging from 0 to 1 in the model application. $\boldsymbol{\theta} = \{\nu_i, \alpha_i, \beta_i, \gamma_i, \sigma_i\}$ is summarized as parameter sets. It further enables to write the likelihood function or probability density function (PDF) from Eqs. (1) and (2) conditional on $\boldsymbol{\theta}$ and Y_i as:

$$f(R|\boldsymbol{\theta}, Y_i) = \frac{\Gamma((\nu_i+1)/2)}{\Gamma(\nu_i/2)} \frac{1}{\sqrt{\nu_i \pi}\, \sigma_i} \left(1 + \frac{1}{\nu_i}\left(\frac{R-(\alpha_i+\beta_i*Y_i+\gamma_i*Z)}{\sigma_i}\right)^2\right)^{-(\nu_i+1)/2} \tag{3}$$

According to the Bayes's theorem (Gelman et al., 2013), the posterior distribution of parameter sets $\boldsymbol{\theta}$ given GR and SPE data, and the prior distribution of parameters $f(\boldsymbol{\theta})$ can be expressed as:

$$f(\boldsymbol{\theta}|\mathrm{R}, Y_i) \propto f(R|\boldsymbol{\theta}, Y_i) f(\boldsymbol{\theta}) \tag{4}$$

The estimation of the posterior distribution $f(\boldsymbol{\theta}|\mathrm{R}, Y_i)$ in Eq. (4) is challenging as its dimension grows with the number of parameters (Renard, 2011). Here, the Markov Chain Monte Carlo (MCMC) technique complied in the Stan programming language is used to address this issue (Gelman et al., 2013). Given that the assumption of the weakly informative priors ensures the Bayesian inferences in an appropriate range (Ma et al., 2020), the priors of $f(\boldsymbol{\theta})$ are initialized as uniform distribution with $\alpha_i, \beta_i, \gamma_i$ at real numbers in Eq. (5), and with ν_i, σ_i at a lower-bound zero of real numbers in Eq. (6).

$$\alpha_i, \beta_i, \gamma_i \sim Uniform(-\infty, +\infty) \tag{5}$$

$$\nu_i, \sigma_i \sim Uniform(0, +\infty) \tag{6}$$

Based on the estimated parameter sets $\boldsymbol{\theta}$ above, the next step is to calculate the bias-corrected SPE R^* at any new site. It can be quantitatively simulated from its posterior distribution in Eq. (7) using the original SPE Y_i^*, and training data R, Y_i:

$$f(R^*|Y_i^*, R, Y_i) = \int f(R^*, \boldsymbol{\theta}|Y_i^*, R, Y_i)\, d\boldsymbol{\theta} \tag{7}$$

Following the rule of joint probabilistic distributions, the right term inside the integral of Eq. (7) is written as:

$$f(R^*, \boldsymbol{\theta}|Y_i^*, R, Y_i) = f(R^*|Y_i^*, R, Y_i, \boldsymbol{\theta}) f(\boldsymbol{\theta}|Y_i^*, R, Y_i) \tag{8}$$

Given that Y_i^* is independent with R and Y_i , the first term of the right side in Eq. (8) is transformed as:

$$f(R^*|Y_i^*, R, Y_i, \boldsymbol{\theta}) = f(R^*|Y_i^*, \boldsymbol{\theta}) \tag{9}$$

Since the parameters $\boldsymbol{\theta}$ are dependent upon the training data R, Y_i, the second term of the right side in Eq. (8) is expressed as:

$$f(\boldsymbol{\theta}|Y_i^*, R, Y_i) = f(\boldsymbol{\theta}|R, Y_i) \tag{10}$$

Therefore, the posterior predictive distribution of R^* in Eq. (7) is written below:

$$f(R^*|Y_i^*, R, Y_i) = \int f(R^*|Y_i^*, \boldsymbol{\theta}) f(\boldsymbol{\theta}|R, Y_i)\, d\boldsymbol{\theta} \tag{11}$$

Since there is no general way to calculate the associated integral in Eq. (11), it is performed again using the MCMC iterations. A numerical algorithm is suggested below: n_{sim} is assumed as the replicates of the post-convergence MCMC samples, and the predicted samples for R^* in Eq. (11) is iterated (i = 1, …, n_{sim}) as follows:

1) Calculate the model parameters $\boldsymbol{\theta}$ from Eqs. (1) to (6) described above;

2) Compute the mean parameter μ_i^* from the regression model of Eq. (2), i.e., $\mu_i^* = \alpha_i + \beta_i * Y_i^* + \gamma_i * Z^*$;

3) Generate the derived quantity from the posterior distribution of R^* in Eq. (11).

3.3 Stage 2: Data merging

On the basis of Stage 1 in Section 3.2, the median value of the posterior samples is used as the bias-corrected SPE. Here, we redefine the bias-corrected SPE as $Y_i'(i = 1,2, \dots, p)$. The formulas of blending the bias-adjusted SPE are shown below:

$$B = \sum_{i=1}^{p} Y_i' * w_i + \varepsilon \tag{12}$$

$$\sum_{i=1}^{p} w_i = 1 \tag{13}$$

$$\varepsilon \sim Normal(0, \sigma) \tag{14}$$

$$w_i \sim Uniform(0,1), i = 1, \dots, p \tag{15}$$

$$\sigma \sim Uniform(0, +\infty) \tag{16}$$

where B means the blended SPE; w_i (i=*1,2,…,p*) stands for the relative weight of the i^{th} bias-corrected SPE with its value ranging from 0 to 1; ε is the residual error with its value at positive real number. Ideally, the blended SPE at the training site *s* and time *t* are close to GR, i.e., *R*(*s*, *t*). Thereby, model parameters $\boldsymbol{\delta}$, including $w_i(i = 1,2, .. p)$ and σ

will be estimated based on GR and bias-corrected SPE at the training sites. With regard to the conditional distribution of blended SPE on the bias-corrected SPE, we propose a Gaussian distribution for residual error modelling. The corresponding PDF is written as follows:

$$f(B|\boldsymbol{\delta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(\frac{B-\sum_{i=1}^{p} Y_i' * w_i}{\sigma})^2) \quad (17)$$

The calculation process of $\boldsymbol{\delta}$ is similar with the parameter estimation described in Stage 1. After the parameters $\boldsymbol{\delta}$ are estimated, similar to Eqs. (7) to (11), the blended SPE at any site and time t can be derived with the bias-corrected SPE and corresponding weights using the MCMC iterations. Finally, we can obtain spatial patterns of blended SPE in terms of the median, standard deviation (SD) and associated credible intervals (e.g., 5% and 95% quantiles) in regions of interest.